Volume 12 Number 1 1984 Nucleic Acids Research

Microcomputer programs for back translation of protein to DNA sequences and analysis of ambiguous DNA sequences

David W.Mount and Bruce Conrad

Departments of Molecular and Medical Microbiology and Biochemistry, College of Medicine, University of Arizona, Tucson, AZ 85724, USA

Received 15 August 1983

ABSTRACT

Three computer programs are described which may be used to translate a DNA sequence into a protein sequence, back translate the protein sequence into an ambiguous DNA sequence, and then do pattern searching in the ambiguous sequence. The programs are written in the C programming language, have been compiled to run on a microcomputer under the CP/M 80 operating system, and may be copied in binary format through a modem. They are also to become available for the IBM/PC.

INTRODUCTION

A number of programs have been described which allow translation of DNA sequences into proteins (1-6). This procedure is useful for finding open reading frames and intervening sequences (7) in long nucleic acid sequences and for predicting peptide subsequences within the complete polypeptide specified by a gene. The peptide can then be synthesized in order to raise antibodies against the *in vivo* gene product. We were interested in back translating protein sequences into an ambiguous sequence representing use of several different codons at each position in the DNA sequence such that the ambigous sequence would specify the same translation product. The usefulness of this approach is three-fold: first, it should be possible to introduce restriction sites into a gene without changing the product; second, it should also be possible to change codon useage to one more favorable for the host organism; third, it should be possible to make changes is the sequence to favor ribosome binding to the resultant mRNA or mRNA stability.

DESCRIPTION OF PROGRAMS

Three programs which act upon a DNA sequence are described: first, one called protein which translates a DNA sequence into an amino acid sequence; a second called backtrans which reverts the amino acid sequence back to an ambiguous DNA sequence; and a third called match which can perform pattern matching on the resultant ambiguous DNA sequence.

Availability of programs

These programs are written in the C programming language, and have been compiled to run on a microcomputer under the CP/M 80 operating system. They should run on any Z80 or 8080 computer with CP/M, including an Apple with the Z80 card. They have been downloaded into a binary format for transmission through the telephone (see accompanying paper by Mount and Conrad for more details).

Protein

The data input for the program protein is a DNA sequence disk file containing the standard 4 bases (A,G,C,T) up to 8000 long, with lin s of any length and spaces and tabs placed anywhere in the file. Lines with any other character are ignored, as are lines with a semicolon in the first column. Any range in the sequence may be specified or the entire sequence. The program output is a disk file of the transaltion product in 3 letter amino acid code with the option of showing the translated codon on an extra line above each amino acid. Termination codons appear as a TER in the output file. The program gives comments at the beginning of the file for later reference. Comments include the name of the original DNA sequence file used and the base range specified and also indicate if there is more than one termination codon. An example of the program acting on a sample DNA sequence file, (Figure 1), to produce a translated sequence (Figure 2) is shown.

Backtrans

Backtrans reverse translates an amino acid sequence disk file given in 3 letter code, capitols or lower case letters, any length lines, with any number of intervening spaces and tabs, into a linear DNA sequence on the computer disk. The output of the program protein described above may be used as the data input file for this program. Backtrans puts in comments in the sequence file after semicolons as a reminder of the source of the protein file. We allow 3 reverse translation options in order to accomodate the 3 stop codons and the 6 possible codons for leu, ser and arg. Please note the amibiguous base assignments e.g M is an A or a C, S a G or a C, etc. shown below. These follow a standard convention which has been proposed (9), and do not conform to those used in the Staden programs (8). A sample program output is shown in Figure 3, following input of data from the disk file which is printed in Figure 2. Utilizing the programs protein and backtrans, it is possible to generate an ambiguous DNA sequence which translates to the same amino acid sequence. However, certain ambiguities will inescapably change the translation product. The program options are shown below:

```
;LEXA GENE FROM MET TO TERMINATION CODON
ATGAAAGCGTTAACGGCCAGGCAACAAGAGGTGTTTGATCTCATCCGTGATCACATCAGCC
AGACAGGTATGCCGCCGACGCGTGCGGAAATCGCGCAGCGTTTGGGGTTCCGTTCCCCAAA
CGCGGCTGAAGAACATCTGAAGGCGCTGGCACGCAAAGGCGTTATTGAAATTGTTTCCGGC
GCATCACGCGGGATTCGTCTGTTGCAGGAAGAGGAAGAAGGGTTGCCGCTGGTAGGTCGTG
TGGCTGCCGGTGAACCACTTCTGGCGCAACAGCATATTGAAGGTCATTATCAGGTCGATCC
TTCCTTATTCAAGCCGAATGCTGATTTCCTGCTGCGCGTCAGCGGGATGTCGATGAAAGAT
ATCGGCATTATGGATGGTGACTTGCTGGCAGTGCATAAAACTCAGGATGTACGTAACGGTC
AGGTCGTTGTCGCACGTATTGATGACGAAGTTACCGTTAAGCGCCTGAAAAAACAGGGCAA
TAAAGTCGAACTGTTGCCAGAAAATAGCGAGTTTAAACCAATTGTCGTTGACCTTCGTCAG
CAGAGCTTCACCATTGAAGGGCTGGCGGTTGGGGTTATTCGCAACGGCGACTGGCTGTAA1
```

Figure 1. Starting test sequence called 'lexa.seq' for forward and reverse translation

```
;translation of sequence file 'lexgene.seq'
;base range 1-609
MET Lys Ala Leu Thr Ala Arg Gln Gln Glu Val Phe Asp Leu Ile Arg
Asp His Ile Ser Gln Thr Gly MET Pro Pro Thr Arg Ala Glu Ile Ala
Gln Arg Leu Gly Phe Arg Ser Pro Asn Ala Ala Glu Glu His Leu Lys
Ala Leu Ala Arg Lys Gly Val Ile Glu Ile Val Ser Gly Ala Ser Arg
Gly Ile Arg Leu Leu Gln Glu Glu Glu Glu Gly Leu Pro Leu Val Gly
Arg Val Ala Ala Gly Glu Pro Leu Leu Ala Gln Gln His Ile Glu Gly
His Tyr Gln Val Asp Pro Ser Leu Phe Lys Pro Asn Ala Asp Phe Leu
Leu Arg Val Ser Gly MET Ser MET Lys Asp Ile Gly Ile MET Asp Gly
Asp Leu Leu Ala Val His Lys Thr Gln Asp Val Arg Asn Gly Gln Val
Val Val Ala Arg Ile Asp Asp Glu Val Thr Val Lys Arg Leu Lys Lys
Gln Gly Asn Lys Val Glu Leu Leu Pro Glu Asn Ser Glu Phe Lys Pro
Ile Val Val Asp Leu Arg Gln Gln Ser Phe Thr Ile Glu Gly Leu Ala
Val Gly Val Ile Arg Asn Gly Asp Trp Leu TER
;NOTE: 1 termination & 5 AUG codons in file
```

Figure 2. Translation of the test sequence called 'lexa.seq' into a disk file called 'lexa.pro' using the program called 'protein. com'

```
;reverse translation of sequence file 'lexa.pro'
;CAUTION-YTN=Leu/Phe,WSN=Ser/Arg,MGN=Arg/Ser,TPP=TER/Trp
;Sequence does not give unique translation product at the above codons
ATG AAP GCN YTN ACN GCN MGN CAP CAP GAP GTN TTY GAY YTN ATH MGN
GAY CAY ATH WSN CAP ACN GGN ATG CCN CCN ACN MGN GCN GAP ATH GCN
CAP MGN YTN GGN TTY MGN WSN CCN AAY GCN GCN GAP GAP CAY YTN AAP
GCN YTN GCN MGN AAP GGN GTN ATH GAP ATH GTN WSN GGN GCN WSN MGN
GGN ATH MGN YTN YTN CAP GAP GAP GAP GAP GGN YTN CCN YTN GTN GGN
MGN GTN GCN GCN GGN GAP CCN YTN YTN GCN CAP CAP CAY ATH GAP GGN
CAY TAY CAP GTN GAY CCN WSN YTN TTY AAP CCN AAY GCN GAY TTY YTN
YTN MGN GTN WSN GGN ATG WSN ATG AAP GAY ATH GGN ATH ATG GAY GGN
GAY YTN YTN GCN GTN CAY AAP ACN CAP GAY GTN MGN AAY GGN CAP GTN
GTN GTN GCN MGN ATH GAY GAY GAP GTN ACN GTN AAP MGN YTN AAP AAP
CAP GGN AAY AAP GTN GAP YTN YTN CCN GAP AAY WSN GAP TTY AAP CCN
ATH GTN GTN GAY YTN MGN CAP CAP WSN TTY ACN ATH GAP GGN YTN GCN
GTN GGN GTN ATH MGN AAY GGN GAY TGG YTN TPP  1
```

Figure 3. Reverse translation of disk file 'lexa.pro' into another disk file called 'lexa.rev' using the program called 'revtrans.com'

```
Sequence file: lexa.rev       Range: 1-300

        10        20        30        40        50
ATGAAPGCNYTNACNGCNMGNCAPCAPGAPGTNTTYGAYYTNATHMGNGA
    HinDIII(5'AGCT)


        60        70        80        90       100
YCAYATHWSNCAPACNGGNATGCCNCCNACNMGNGCNGAPATHGCNCAPM
                             SmaI(FCCC)          HinDIII(5'AGCT)

       110       120       130       140       150
GNYTNGGNTTYMGNWSNCCNAAYGCNGCNGAPGAPCAYYTNAAPGCNYTN
                        PstI(TGCA3')       HinDIII(5'AGCT)

       160       170       180       190       200
GCNMGNAAPGGNGTNATHGAPATHGTNWSNGGNGCNWSNMGNGGNATHMG
                                   PstI(TGCA3')
                                     SmaI(FCCC)
                                           BamHI(5'GATC)
                                           EcoRI(5'AATT)
                                               HinDIII(5'AGCT)

       210       220       230       240       250
NYTNYTNCAPGAPGAPGAPGAPGGNYTNCCNYTNGTNGGNMGNGTNGCNG
    PstI(TGCA3')                                PstI(TGCA3')

       260       270       280       290       300
CNGGNGAPCCNYTNYTNGCNCAPCAPCAYATHGAPGGNCAYTAYCAPGTN
                                                 BamHI(5'GATC)


Pattern identifier    Pattern matched     Base number matched
------------------    ---------------     -------------------

SmaI(FCCC)            CCCGGG               80, 188
PstI(TGCA3')          CTGCAG              125, 186, 205, 248
EcoRI(5'AATT)         GAATTC              194
BamHI(5'GATC)         GGATCC              194, 300
HinDIII(5'AGCT)       AAGCTT                5,  99, 143, 198
```

Figure 4. Pattern matching of list of restriction sites to the sequence called 'lexa.rev' shown in Fig. 3 using the program called 'match' which recognizes ambiguous base sequences.

1. TTP for Leu, AGY for Ser, AGP for Arg, TAP for TER.
2. CTN for Leu, TCN for Ser, CGN for Arg, TGA for TER.
3. YTN for Leu, WSN for Ser, MGN for Arg, TPP for TER.

where P=A/G Y=C/T N=A/G/C/T W=A/T S=G/C M=A/C

NOTE: Option 3 can give an altered translation product

Match

The program match is an extension of one called resenz, which we have described before (6). These programs both read patterns and their names (e.g. EcoRI GAATTC) from a match table in a disk file, or accept input patterns from the terminal. We have expanded the pattern matching capabilities of resenz to include ambiguous base letters which represent all possible combinations of bases (9), and are also used in the program backtrans described above. A '+' representing any of the 4 DNA bases or no base at the position is also recognized. This program will accept as its data input an unambiguous or ambiguous DNA sequence, such as shown in Figures 1 and 3, respectively. Given input data from the ambiguous DNA sequence file shown in Figure 3, and searching for patterns that would be recognized by SmaI, PstI, EcoRI, BamHI and HindIII, the program was able to find the matches shown in the first 250 base pairs of the sequence shown in Figure 4.

ACKNOWLEDGEMENTS

This laboratory is supported by grants from the National Science Foundation and the National Institutes of Health.

REFERENCES

1. Staden, R. (1977) Nucl. Acids Res. 4, 4037-4051
2. Korn, L.J., Queen, C.L., and Wegman, M.N. (1977) Proc. Natl. Acad. Sci. USA 74, 4401-4405.
3. Gingeras, T.K., Roberts, R.J. (1980) Science 209, 1322-1328.
4. Sege, R., Soll, D., Ruddle, F.H., Queen, C. (1982) Nucl. Acids Res. 9, 437-443
5. Blumenthal, R.M., Rice, P.J., Roberts, R.J. (1982) Nucl. Acids Res. 10, 91-114
6, Conrad, B., Mount, D.W. (1982) Nucl. Acids Res. 10, 31-37
7. Staden, R., McLaclan, A.D. (1982) Nucl. Acids Res. 10, 141-156
8. Staden, R. (1979) Nucl. Acids Res. 6, 2601-2610
9. Lathe, R. (1983) Nomenclature for incompletely specified bases in nucleic acid sequences - IUB-IUPAC Joint Commission on Biochemical Nomenclature.